**Generalized Model of Mixtures with Feature Selection: An Embedded Approach**

**Abstract**

Clustering is a commonly used technique to discover patterns in the dataset and most traditional clustering algorithms focus on datasets that contain either continuous or categorical variables. Since real applications often have mixed datatypes, mixture model has attracted great attention lately due to its potentials of handling mixed data. Mixture model is a “soft assignment” method which assumes that the instances are generated from a mixture of underlying probability distributions. Expectation Maximization (EM) algorithm is applied to estimate the model parameters. A known challenge for mixture model is high-dimensional dataset where the parameter estimation is computational costly. Noting some features may not make significant contributions for the clustering, we develop an embedded feature selection algorithm for generalized model of mixtures by adding a feature selection step (S) to the EM procedure. Specifically, we introduce a feature index (*FI*), a metric indicating the probability of assigning a data point to a certain clustering group. The *FI* index reveals the contribution of the feature in the clustering process thus can assist the feature selection. We conduct theoretical analysis to justify the use of *FI* for feature selection. In addition, to demonstrate the efficacy of the proposed ESM, one synthetic dataset, one benchmark dataset and a real world Alzheimer’s Disease dataset are studied.

Key Words: Clustering; Generalized Model of Mixture; Mixed data; feature selection

**1. Introduction**

Clustering is an unsupervised data mining technique to group the data into different segments by discovering the patterns within the dataset [1]. The underlying assumption is the data points from the same group are more like each other than the data from other groups. Due to the nature of the problem, clustering has been broadly applied to many fields including manufacturing [2], biology [3], finance [4] and astronomy [5], just to name a few.

Most existing clustering algorithms are designed to tackle single data type (e.g., continuous, categorical) thus limit its applications to the real-world problems which often contain mixed data types. The intuitive solution to a mixture dataset is to convert the data set to a single type by either transforming the categorical features to numbers or converting the continuous features into categorical features. Classic clustering methods can then be applied afterwards. One example is to dummy coding all categorical features to continuous features [6]. There are three issues associated with this. First, the dimension of data set is increased and this may cause problems when the number of categorical features or the level of the categorical feature is large. Second criticism is the semantic similarity in the original data set may be lost during the transformation [7]. The third issue is it is non-trivial to give correct numeric values to categorical values like color [8]. While converting continuous features to categorical ones may be less problematic, the discretization process may again lose information [8]. An alternative approach on mixture dataset is to define new distance measures and cost function designed specifically for the types of the data. Enormous efforts have been invested on improving k-means and k-prototypes clustering algorithm. For example, based on k-prototypes, new measures such as Gower’s distance [9] are introduced to calculate the dissimilarity between data objects and prototypes of clusters. Gower’s distance usually involves weights specified for the distance of continuous features and the distance of categorical features. As criticized by Foss et al. (2016), determining the weights is critical for the clustering and yet there is currently no explicit guideline on how to assign the weights for optimal clustering outcomes. The third approach on mixed data takes ensembled methods [10]. The idea is simple. The mixed data set is first divided into two sub-datasets: the categorical dataset and the continuous dataset. Traditional clustering algorithms designed for different types of datasets are applied respectively and the clustering results on the two sub-datasets are combined via a sequential combination method. The main issue of the ensemble method is that the clustering algorithms are biased by partial of the dataset. And, using single type of data does not take advantage of the complementary information from other data.

The three approaches reviewed above are mostly instance-based focusing on the data points instead of the dataset distributions. One may argue that converting the categorical features into continuous features from the first approach does use the normal distribution as the guideline. However, this is under the assumption that the categorical data can be represented as continuous data from a normal distribution. The true distribution on the original categorical data is not utilized. In contrast, model-based methods look into the probabilistic distribution from the true data. It assumes that the instances (data points) are generated from a mixture of underlying probability distribution. Literature terms this approach as model-based clustering (Banfield & Raftery 1993; Bensmail et al. 1997; Fraley & Raftery 1998a, 1998b), mixture likelihood clustering (McLachlan & Basford 1988; Everitt 1993), mixture-model clustering (Jorgensen & Hunt 1996; McLachlan, et al. 1999) and Latent Class cluster analysis (Vermunt & Mgidson, 1996). In general, model-based approaches (a.k.a. Latent Class models) have several advantages in clustering mixed type data. (1) Statistical metrics (e.g., mean and variance) derived from the data distribution instead of the distance between the data points are used in the modeling. The model uses the probability to describe the degree of the data point belonging to the cluster and the probability is updated iteratively during the clustering. Since most real world problems are uncertain by nature, the use of this “soft assignment” approach may be a better alternative comparing to the “hard assignment” (e.g., k-means) [11][12]. (2) Latent Class model enjoys the flexibility in choosing the distribution forms for each component. The continuous features can be modeled as normal distribution while the categorical features can be estimated from multinomial or Poisson distribution. In addition, some restrictions can be imposed on the model parameters to simplify the model structure and avoid overfitting. For example, the covariance matrix can be restricted to be diagonal for high dimensional data to reduce the number of parameters. Formal statistical test can also be applied to check the validity of parsimonious model [12] [13]. (3) The third advantage of the Latent Class model is it is scale-free, that is, the clustering results are independent from the data being normalized or not. For distance based clustering algorithms like k-means, the scaling has been one of the main criticisms. Especially when handing mixed type of data, the categorical feature may influence the scaling of continuous features [1]. (4) Latent Class models have formal criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) to decide the number of clusters [14].

While the Latent Class model may have great potentials in handling mixed type of data, it faces some challenges for high-dimensional dataset. It is noted that among the large number of features, some may not truly contribute to delineate the cluster profiles. Inclusion of these “noisy features” requires more parameter estimations, which cost unnecessary computational overhead. Additionally, the noisy features will confuse the model to identify the true structure of the clusters [15]. Recognizing the issue, in this research, we first develop a Generalized Model of Mixtures (GMoM) where a multivariate normal distribution is employed to describe continuous features, Bernoulli distribution and multinomial distribution are used to model binary and nominal features respectively. Next, we propose a novel Feature Index (*FI*) based on Kullback-Leibler (KL) divergence. *FI* is a measure based on the posterior probability of assigning data points to the cluster groups. If the inclusion and exclusion of a feature show no significant difference on the *FI* measure, we conclude this feature is not truly contributing to the clustering. Thus, *FI* can be used to rank and select important features for clustering and reduce the dimensionality of the feature space. One unique advantage of our proposed approach is the assessment of *FI* can be naturally embedded in the model parameters estimation procedure. The parameter estimation procedure for mixture models such as Expectation Maximization (EM) are often iterative and can be computational expensive. The classical feature selection technique such as wrappers [16] that cycles between selecting feature subsets and estimating model parameters, can become unfeasible in the feature selection for mixture model setting. This embedded approach that simultaneously select features and estimate model parameters can be more computational efficient compared to wrapper methods.

The remainder of this study is organized as follows. Section 2 reviews the basics of Latent Class Model and EM algorithm. Section 3 presents the proposed algorithm in details. Next, three sets of experiments on one synthetic dataset, one benchmark dataset and one real application dataset are illustrated in Section 4. In Section 5, the conclusion and future direction are presented.

**2. Review of Latent Class Model and EM**

Suppose we have a data matrix with *N* data points and *p* features in which the row is . Let be a vector of *p* features where each feature can be continuous, binary or nominal. Let be the value of the sample for the feature.

In the Latent Class model, we assume that the data can be grouped into *K* clusters. For each cluster *k* is an associated probability . The joint distribution of the observed features is a finite mixture of probabilities :

The probability density function is discussed for binary, nominal, and continuous features separately [11]. Specifically,

* For a binary feature, we take the Bernoulli distribution:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

* For nominal features, the indicator feature is replaced by a vector-valued indicator function with its *s*th element being defined as

where denotes the number of categories of feature and . The distribution assumed is multinomial

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where is the probability that an object who is in class k will belong to category s for feature .

* For continuous features, normal distribution is employed for each single continuous feature:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where is the location parameter of the continuous feature in class and is the variance of the feature taken as constant across classes. However, normal distribution neglects the correlations between the continuous features. We extend Latent Class model by relaxing the independence assumptions on the continuous features using multivariate normal distribution. Among the *p* features, let the first *c* features are continuous, we have:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

where is the mean vector of the continuous feature in class and is the covariance matrix.

Then the joint likelihood of sample belonging to group is

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Note the correlations among continuous features can be well represented by multivariate normal distribution. For mixed-type data, literature shows existing methods typically adopt a normal-multinomial finite mixture [17]–[21]. Location model [22] can be employed to allow a distinct distribution for the continuous variables for each unique combination of categorical levels. This approach accounts for any possible dependence structure between continuous and categorical variables, however, it becomes infeasible when the number of categorical variables or number of levels within each categorical variable is large [7]. In addition, deriving all possible dependence structures indeed is to “blend” the categorical features into the structure. As a result, the true identity of the categorical features may be lost which in turn, will impede the feature selection process. Therefore, in this research, we relax the dependence assumptions on the categorical features (binary and nominal) and assume the categorical features are independent to each other.

Given the probability density function for binary, nominal, and continuous features, Latent Class model aims to the maximize the log-likelihood of the full data for all features, which is shown in Equation (6).

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

EM is a commonly used four-step algorithm to estimate the parameters including mixing coefficients and parameters for each distribution, e.g., mean and covariance for normal distribution. The algorithm starts from initializing parameters, and evaluates the initial value of log likelihood function. In the second step (known as E step), the EM evaluates the responsibilities under the current parameter settings. The responsibility is defined as the probability of assigning a data point to a specific clustering group:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

In the third step, the EM re-estimates the parameters given the responsibilities. The estimation method is maximum likelihood, and therefore the third step is called “M step”. The final step is check the convergence of log-likelihood. If the log likelihood difference between two iterations is small enough (e.g., less than a small number), it is converged. Otherwise, the algorithm goes back to the E step initiating the next iteration. For details of the EM algorithm, interested readers are referred to [23]. In the next section, we present the Feature Index used for feature selection on this Generalized Model of Mixtures embedded in the EM algorithm.

**3. Proposed Method**

**3.1. Feature Index**

Our proposed Feature Index (*FI*) takes advantage of the responsibility measures in the E step. Let us consider the responsibilities , the probability of assigning the data point to cluster *k*, if we remove one specific feature, responsibilities shall change. Specifically, let the full feature space with *p* features be , the feature space excluding feature be . Here, we denote the responsibility on the full feature space as and the responsibility on the reduced feature space (excluding feature) as which is related to the data point and the cluster.

To capture the difference between and , we use the Kullback-Leibler, or simply KL divergence [24]. The KL divergence is closely related to relative entropy, information divergence, and information for discrimination. It is a non-symmetric measure of the distance between two probability distributions and . The KL divergence of from is defined as

which measures the information lost when is used to approximate . The reason of choosing KL divergence instead of other distance measure such as Euclidian metric is that KL divergence has a statistical meaning, which is helpful in handling probability distributions. The statistical properties are further explained in the theoretical analysis section.

Based on the definition of KL divergence, the proposed *FI* is defined as the KL divergence between two responsibilities averaged over *N* data points and *K* clustering groups. Since the two responsibilities are calculated under the full feature set and the feature set excluding the feature respectively, the divergence between the two responsibilities reveals the importance of feature to the clustering. It is written as:

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

The assumption behind our proposed method is that if is smaller than a pre-defined threshold, the contribution of feature in deciding the assignments of data points to clusters is trivial thus can be neglected. The feature can be removed during the feature selection process under the condition that converges over the iterations. Concerning the convergence criteria for , we evaluate the changes of between the current and previous iteration, let say, if it is less than a small number, e.g., 0.0005, we conclude converges. Regarding the pre-defined threshold for feature selection, it can be set based on the approximate number of features to be selected (see experiments in Section 4 for details). Table 1 summarizes the ESM algorithm with the proposed S step highlighted.

Table 1 Pseudo Code on Extending EM with Feature Index on Generalized Model of Mixtures

1. Initialize the parameters which is from a converged Latent Class model.
2. **E step.**  Evaluate the responsibilities using the current parameter values

|  |  |  |
| --- | --- | --- |
|  |  |  |

and responsibilities after excluding each feature

1. **S step.** Calculate the difference between responsibilities before and after excluding feature at iteration t.

If (converged) and is small enough, then discard the feature with smallest and update the full feature space *F*.

1. **M step.** For reduced data with feature space *F*, re-estimate the parameters using the current responsibilities

for binary features:

for nominal features*:*

for continuous features*:*

posterior probability*:*

1. Evaluate the log likelihood

|  |  |  |
| --- | --- | --- |
|  |  |  |

If the parameters or the log likelihood are not converged, go back to step 2.

In the next section, we provide theoretical analysis on the statistical properties of *FI*.

**3.2 Theoretical Analysis on Feature Index**

The proposed *FI* is the KL divergence between two responsibilities averaged over *N* data points and *K* clustering groups. In the following equation, function is defined as in equation (1), (2) and (4) for bernoulli, multinomial and multivariate normal distribution respectively. The *FI* is calculated by Equation (8). We decompose the *FI* into two parts as shown in Equation (9) and (10). Note that in the following decomposition, the continuous features are the first c out of p features as used in Equation (4) before.

If the feature is binary or nominal,

If the feature is continuous, let be the corresponding vector of after excluding the feature.

Using the conditional probability formula, ,

From equations (9) and (10), the *FI* is decomposed into two: (1) weighted conditional log-likelihood of feature *j* given all other features; (2) weighted log-likelihood ratio for all features vs. all except feature *j*. The first term reveals the absolute contribution of feature *j* to the model and the second term can be considered as the relative contribution of the feature.

Let us delve more into the second terms from equations (9) and (10). For equation (9), the second term is the likelihood ratio shown in Equation (11),

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

In Equation (11), the denominator is the likelihood of the full feature set and the numerator is the likelihood of the feature set excluding feature. For each GROUP , the denominator is the numerator times , the likelihood for the data point of feature on group , which is between 0 and 1. Hence, this second term is greater or equal to 1. Similarly, for Equation (10), the likelihood ratio in the second term as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

In Equation (12), for each , the denominator is the numerator times , which is between 0 and 1. Thus, the likelihood ratio in Equation (10) is also greater or equal to 1.

From Equation (11) and (12), we obtain the conclusion that the second term of increases as the likelihood difference between the full model and the model excluding feature increases. In special case, when the likelihood of the full model is the same as the model excluding one feature, then the second term is equal to one.

From the analysis, we conclude *FI* is an integrated measure of the absolute and relative contribution of a given feature. A feature deemed to be important to the clustering model can be revealed from two aspects: a large likelihood showing dominating absolute contribution and/or a large log-likelihood ratio, indicating relative contribution to the full feature set.

In the next section, we will use some experiments on various datasets including synthetic, benchmark and real application to validate the efficacy of *FI*.

**4. Experiments**

In the section, one synthetic dataset, one benchmark dataset and one medical application dataset are studied to demonstrate the performance of proposed algorithm. Since the ground truth of all the datasets are known, we use the following two metrics for the performance evaluation: (1) RFS: the percentage of relevant features being selected; (2) Accuracy: the percentage of instances correctly clustered.

**4.1. Experiments on Synthetic Datasets**

We design the synthetic dataset with four relevant features, and two clusters. Additional irrelevant features are generated as “noise features”. In this experiment, we include 15 features with , (continuous) and (categorical) being the relevant features for clustering and remaining 11 features being irrelevant features. The continuous relevant features are generated from two-component mixture of Gaussian distributions with 300 data points for each component. The categorical relevant features are generated from multinomial distributions for each component. In addition, we purposely add correlations between the features including ~ , ~ and ~ correlations. The experiment setting is summarized in Table 2. Similarly, 10 runs are conducted for each experiment.

Table 2. Experiment setting for synthetic dataset

|  |  |
| --- | --- |
| # of features | 10 continuous (2 relevant +8 irrelevant)  +5 categorical (2 relevant+3 irrelevant) |
| # of groups/components | a mixture of 2 components |
| # of data points | 600 |
| # of repeated runs | 10 |
| Distribution of relevant features for each component (continuous) |  |
| Distribution of irrelevant features (continuous) |  |
| Distribution of relevant feature (categorical) | Component 1: )  Component 2: ) |
| Distribution of relevant feature (categorical) | Component 1: )  Component 2: ) |
| Distribution of irrelevant feature  (categorical) | )  )  ) |
| # of repeated runs | 10 |

In the experiment, we first test the classical Latent Class Model on the simulated data. The accuracy is 82.83%. Then we test our proposed algorithm on the synthetic dataset. That is, we do feature selection on latent class model. The accuracy improves to 94.63%. In addition, the selected features are and, which are exactly the relevant features we used to simulate the clustering group.

Figure 1 shows all relevant features , and have significantly higher *FI* values than the other 11 irrelevant features. We conclude *FI* is able to identify the relevant features with dependencies and exclude the irrelevant “noisy features” resulting much improved clustering performance.

C:\Users\yinlinfu\Dropbox\ESM phase II\paper\figure\figure1_1.tiff

Figure 1. for each feature over iterations on 600 data points. The large values of for , and in the final iteration show that and are critical for clustering.

**4.2. Benchmark Data**

In this section, we compare our proposed algorithm with other clustering approaches that can handle mixed data on benchmark dataset obtained from UCI Machine Learning Repository [30]. The studied dataset is the heart rate disease dataset provided by Cleveland Clinic. The dataset has 303 instances with 6 numeric and 8 categorical features. The instances are labelled as two classes: healthy or sick (with heart disease).

Table 3. Summary of features for Heart Disease data.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Notation | Mean ± SE | Category |
| Age |  | 54.4 ± 9.1 | Continuous |
| Sex |  | 0,1 | Binary |
| Chest pain type |  | 1,2,3,4 | Nominal |
| Resting blood pressure |  | 131.3 ± 17.9 | Continuous |
| Serum cholestoral in mg/dl |  | 249.7 ± 51.7 | Continuous |
| Fasting blood sugar>120 mg/dl |  | 0,1 | Binary |
| Resting electrocardiographic results |  | 0,1,2 | Nominal |
| Maximum heart rate achieved |  | 149.7 ± 23.2 | Continuous |
| Exercise induced angina |  | 0,1 | Binary |
| Oldpeak=ST depression induced by exercise relative to rest |  | 1.1 ± 1.1 | Continuous |
| The slope of the peak exercise ST segment |  | 1,2,3 | Nominal |
| Number of major vessels colored by flourosopy |  | 0,1,2,3 | Nominal |
| Thal: 3=normal; 6=fixed defect; 7=reversible defect |  | 3,6,7 | Nominal |

In literature, there are four other clustering approaches reporting accuracy on the heart disease data. The K-prototype and K-medoids are two classical clustering approaches that can handle mixed data. The reported best accuracies on heart disease data are 81.0% and 76.5% respectively [25][10]. The third approach is ensemble clustering proposed by Z.He and X.Xu etc. [10]. The general idea of ensemble clustering is to first divide the original mixed dataset into two subsets: the pure categorical and the pure numeric dataset. Then they use the existing clustering algorithms designed for one certain type of dataset to cluster each dataset. Finally, they combine the clustering results on two datasets and generate final clusters. The best accuracy on the heart disease data is 81.3%, which is slightly better than K-prototype. The fourth approach is called UFL Fuzzy ART [25], which extends the basic unsupervised feature learning (UFL) for mixed type data using fuzzy adaptive resonance theory (ART). The best reported accuracy on heart disease data is 81.5%, the highest among all current approaches.

In addition to the four approaches provided in literature, we also test the classical latent class model (LCM) on the heart disease data. The best accuracy obtained by LCM is 78.9%, which is slightly lower than other approaches. Then finally, we employ our proposed feature selection algorithm using feature index on the heart disease data and the best average accuracy is 83.3%, which is higher than any other algorithms listed before. To have a better view of the comparison, the performance of each algorithm is summarized in Figure 2.



Figure 2. Summary of Accuracy performance on Heart disease data.

In addition, we show the importance of each feature by Feature Index. The plot is given below.

C:\Users\yinlinfu\Dropbox\ESM phase II\paper\figure\figure2_2.tiff

Figure 3. for each feature over iterations on heart disease dataset. The large values of for , and in the final iteration show that and are critical for clustering.

**4.2. Real World Application: Alzheimer’s Disease**

Alzheimer’s Disease (AD) is a progressively neurodegenerative disease which is the most frequent type among elderly dementia patients. In the U.S., approximately 5.2 million people over 60 are afflicted by AD (Alzheimer's Association,2008). This drives a great amount of research investigating ways to slow down the AD progression and detect AD at early stage for better treatment or even prevent the disease. Mild cognitive impairment (MCI) is a syndrome defined as cognitive decline greater than expected for individuals during the course of aging but that does not interfere notably with activities of daily life [27]. It is an intermediate stage between normal aging with mild cognitive decline and dementia where cognitive impairment is more severe even impacting daily function. Though it is distinct from dementia, MCI patients with memory complaints and deficits (amnestic mild cognitive impairment) have high risks of progression to AD [27], [28]. The early diagnosis of MCI stage is becoming essential when the interventional strategies may be more effective.

Extensive research has investigated predictive model for AD in hoping to predict the risk of each individual patient converting to AD and this is still on-going effort. The focus of this research is to identify the underlying patient cohort structures which may discover patient subtypes for interventional treatment. In this study, we have collected 317 patients’ data from Alzheimer’s disease neuroimaging initiative [29], a large scale online repository designed to identify more sensitive and accurate methods to detect Alzheimer’s disease at earlier stage and mark its progress via biomarkers. Specifically, the baseline data is collected to evaluate the efficacy of our proposed method in AD early detection. Among all these patients, 22 are AD, 172 are MCI and 123 are Normal Controls (NCs). For each patient, we obtain PET, and MR images and cognitive tests (see Table 4).

Table 4. Summary of features for Alzheimer’s Disease data.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Notation | Mean ± SE | Category |
| Age |  | 72.9 ± 7.3 | Demographic |
| Mini-Mental State Examination (MMSE) |  | 28.2 ± 2.3 | Cognitive Test |
| Clinical Dementia Rating (CDR) Score |  | 1.1 ± 1.6 | Cognitive Test |
| Volume of hippocampus |  | 7157 ± 1160 | MRI biomarkers |
| Volume of ventricles |  | 35086 ± 19333 | MRI biomarkers |
| Whole Brain |  | 1048732 ± 112296 | MRI biomarkers |
| Entorhinal |  | 3677 ± 741 | MRI biomarkers |
| Volume of Intracranial |  | 1514196 ± 156568 | MRI biomarkers |
| Hypometabolic Convergence Index (HCI) |  | 10.9 ± 5.6 | FDG-PET biomarkers |
| Statistical region of interest (sROI) |  | 1.2 ± 0.07 | FDG-PET biomarkers |
| mean cortical Standard Uptake Value Ratio with cerebellum as reference region (mcSUVRcere) |  | 1.1 ± 0.2 | F-AV45-PET biomarkers |
| mean cortical Standard Uptake Value Ratio with corpus callosum and centrum semiovale combined as reference region (mcSUVRwm) |  | 0.77 ± 0.17 | F-AV45-PET biomarkers |
| Gender |  | Two levels  (176:141) | Demographic |
| AV45 |  | Two levels  (202:115) | F-AV45-PET biomarkers |
| APOE |  | Three level  (22:103:192) | Gene information |

Table 5 presents the correlation matrix of the features. Clearly, some features are highly dependent to each other, for example, the clinical test scores MMSE (*f*2) vs. CDR (*f*3), MRI biomarkers volume of hippocampus (*f*4) vs. whole brain (*f*6), FDG-PET biomarkers HCI (*f*9) vs. sROI (*f*10).

Table 5. Correlation Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *f1* | *f2* | *f3* | *f4* | *f5* | *f6* | *f7* | *f8* | *f9* | *f10* | *f11* | *f12* |
| *f1* | 1.00 | -0.09 | -0.02 | -0.37 | 0.41 | -0.34 | -0.15 | 0.02 | 0.03 | -0.35 | 0.12 | 0.19 |
| *f2* |  | 1.00 | **-0.72** | 0.44 | -0.19 | 0.19 | 0.41 | 0.03 | -0.48 | 0.41 | -0.38 | -0.47 |
| *f3* |  |  | 1.00 | -0.41 | 0.22 | -0.14 | -0.42 | 0.02 | **0.55** | -0.45 | 0.38 | 0.48 |
| *f4* |  |  |  | 1.00 | -0.33 | **0.56** | **0.64** | 0.29 | -0.39 | **0.54** | -0.27 | -0.39 |
| *f5* |  |  |  |  | 1.00 | 0.06 | -0.11 | 0.45 | 0.28 | -0.43 | 0.11 | 0.44 |
| *f6* |  |  |  |  |  | 1.00 | 0.49 | **0.79** | 0.00 | 0.35 | -0.07 | -0.05 |
| *f7* |  |  |  |  |  |  | 1.00 | 0.30 | -0.27 | 0.38 | -0.19 | -0.25 |
| *f8* |  |  |  |  |  |  |  | 1.00 | 0.15 | 0.05 | 0.04 | 0.14 |
| *f9* |  |  |  |  |  |  |  |  | 1.00 | **-0.65** | 0.32 | 0.49 |
| *f10* |  |  |  |  |  |  |  |  |  | 1.00 | -0.30 | -0.41 |
| *f11* |  |  |  |  |  |  |  |  |  |  | 1.00 | **0.84** |
| *f12* |  |  |  |  |  |  |  |  |  |  |  | 1.00 |

One challenge in applying classical latent class model to the data is that the correlations between continuous features are not captured. Another challenge is that some noise features degrade the performance of Latent Class model.

Table 6 shows the clustering results from classical Latent class model using all 15 features. The overall accuracy of correctly identifying patients to disease types is only 59.31%. Table 7 shows the clustering results after feature selection on the generalized model of mixtures. The overall accuracy improves from 59.31% to 84.86% compared with original LCM. For the AD cohort, the proposed algorithm clusters 22 out of 22 AD patients correctly (100%). For the NC cohort, our algorithm identifies 114 out of 123 correctly, 1 NC is put into the AD group, and another 8 NCs are labeled as MCI. The accuracy of NC cluster is 92.68%. The results on MCI cohort also improves compared with LCM using full feature set. The feature selection on the GMoM can identify 133 out of 172 MCI correctly (77.33%) with only one mislabeled as NC and 38 mislabelld as AD.

Table 6. Confusion matrix of classical Latent Class model using all features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | GroundTruth | | |  |
|  |  | NC | AD | MCI | Overall accuracy |
| Clustering | NC | 92 | 0 | 88 |  |
| AD | 0 | 21 | 9 |  |
| MCI | 31 | 1 | 75 |  |
|  | Total | 123 | 22 | 172 |  |
|  | Accuracy | 74.80% | 95.45% | 43.60% | **59.31%** |

Table 7. Confusion matrix for clustering after feature selection using Feature Index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | GroundTruth | | |  |
|  |  | NC | AD | MCI | Overall accuracy |
| Clustering | NC | 114 | 0 | 1 |  |
| AD | 1 | 22 | 38 |  |
| MCI | 8 | 0 | 133 |  |
|  | Total | 123 | 22 | 172 |  |
|  | Accuracy | 92.68% | 100% | 77.33% | **84.86%** |

Figure 3 shows the feature index for all 15 features. The three features with highest feature index are: CDR(), HCI () and mcSUVRcere ().

C:\Users\yinlinfu\Dropbox\ESM phase II\paper\figure\figure3_3.tiff

Figure 3. for each feature over iterations on AD data. The large values of for , and in the final iteration show that , and are critical for clustering.

We want to emphasize that the results on MCI are not surprising. Clinically, MCI cohort has subtypes: MCI converter and MCI non-converter. MCI converter refers to the patient positively diagnosed as AD in the follow-up exams. Fortunately, ANDI is a rich data repository with longitudinal data available. We collect the updated patient staging information from the follow-up visit to explore the composition of the MCI group. For illustration purpose, we show a 2D plot of the two most relevant features CDR (*f*3) and mcSUVRcere (*f*11).

C:\Users\yinlinfu\Dropbox\ESM phase II\paper\figure\figure5.tiff

Figure 4. Clustering shown in two feature space: CDR (*f*3) and mcSUVRcere (*f*11).

The points are colored by the real diagnosis results: AD (red), MCI (green) and NC (blue). The shapes signify the converters: triangle for NC converting to MCI, square for NC converting to AD and cross for MCI converting to AD. In Figure 4, 16 out of 26 green crosses are on the boundary between MCI and AD clusters, but close to AD. That is, among the 172 MCIs, 26 are staged as AD in the follow-up visit. Using baseline data, the 17 out of 38 MCIs mislabeled as AD (Table 7) are indeed the converted. The blue triangle represents the patient converting from NC to MCI. In the 7 blue triangles, there is one special point, which is diagnosed as NC in the first visit. However, in clustering, the point is closer to MCI than NC in Figure 4 and mislabeled as MCI in Table 7, which is verified by the diagnosis of MCI in the second visit. Indeed, the clustering technique can help to capture the convert between disease types.

**5. Conclusion and Future Work**

Latent Class Model, as a soft clustering methodology, has attracted great attention due to the distinct advantages from it statistical foundation. However, its performance deteriorates notably if the dataset has many noisy features irrelevant to the clustering process. This research proposes a new metric: Feature Index (*FI*). Traditional EM algorithm for Latent Class modeling parameter estimation is extended with a S step using *FI* for feature selection. Our proposed embedding the feature selection into the EM algorithm preserves the good properties of EM algorithm such as guaranteed convergence and optimum determination of the clustering number. To evaluate the performance of the proposed algorithm, experiments on one synthetic data set, one benchmark dataset and one Alzheimer’s Disease (AD) dataset are conducted. The experiments on synthetic and benchmark dataset show the proposed *FI* is able to identify the relevant features and improved clustering accuracy comparing to classical Latent Class model without feature selection. Other than improved clustering result, experiment on AD indicates the model based clustering with feature selection may potentially identify the patient subtypes which is crucial for patient treatment planning. While promising, the algorithm is limited in handling dependency between categorical features. In the future, we may tackle the issue for the dependence among categorical features and the dependence between categorical group and continuous data group.

**Conflict of Interest**

None declared.

**Acknowledgments**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**References**

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[2] J. Yu, “Fault detection using principal components-based gaussian mixture model for semiconductor manufacturing processes,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 3, pp. 432–444, 2011.

[3] B. Mazoyer *et al.*, “Gaussian Mixture Modeling of Hemispheric Lateralization for Language in a Large Sample of Healthy Individuals Balanced for Handedness,” *PLoS One*, vol. 9, no. 6, p. e101165, Jun. 2014.

[4] A. Lindemann, C. L. Dunis, and P. Lisboa, “Probability distributions, trading strategies and leverage: An application of Gaussian mixture models,” *J. Forecast.*, vol. 23, no. 8, pp. 559–585, 2004.

[5] K. J. Lee, L. Guillemot, Y. L. Yue, M. Kramer, and D. J. Champion, “Application of the Gaussian mixture model in pulsar astronomy,” *Mon. Not. R. Astron. Soc.*, vol. 424, no. 4, pp. 2832–2840, 2012.

[6] C. Hennig and T. F. Liao, “How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification,” *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 62, no. 3, pp. 309–369, 2013.

[7] A. Foss, M. Markatou, B. Ray, and A. Heching, “A semiparametric method for clustering mixed data,” *Mach. Learn.*, vol. 105, no. 3, pp. 419–458, Dec. 2016.

[8] A. Ahmad and L. Dey, “A k-mean clustering algorithm for mixed numeric and categorical data,” *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.

[9] J. C. Gower, “A General Coefficient of Similarity and Some of Its Properties,” *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.

[10] Z. He, X. Xu, and S. Deng, “Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach,” p. 14, 2005.

[11] I. Moustaki and I. Papageorgiou, “Latent class models for mixed variables with applications in Archaeometry,” *Comput. Stat. Data Anal.*, vol. 48, no. 3, pp. 659–675, 2005.

[12] C. Fraley and a E. Raftery, “How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis,” *Comput. J.*, vol. 41, no. 8, pp. 578–588, 1998.

[13] D. B. Dahl, “Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model,” *Bayesian inference gene Expr. proteomics*, pp. 201–218, 2006.

[14] A. E. Raftery and N. Dean, “Variable Selection for Model-Based Clustering,” *J. Am. Stat. Assoc.*, vol. 101, no. 473, pp. 168–178, 2006.

[15] C. Maugis, G. Celeux, and M. L. Martin-Magniette, “Variable Selection for Clustering with Gaussian Mixture Models,” *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[16] S. Adams and P. A. Beling, “A survey of feature selection methods for Gaussian mixture models and hidden Markov models,” *Artif. Intell. Rev.*, pp. 1–41, 2017.

[17] C. J. Lawrence and W. J. Krzanowski, “Mixture separation for mixed-mode data,” *Stat. Comput.*, vol. 6, no. 1, pp. 85–92, Mar. 1996.

[18] L. A. H. Murray A. Jorgensen, “Mixture Model Clustering of Data Sets with Categorical and Continuous Variables,” pp. 278–283, 1996.

[19] G. J. McLachlan, R. W. Bean, and D. Peel, “A mixture model-based approach to the clustering of microarray expression data,” *Bioinformatics*, vol. 18, no. 3, pp. 413–422, Mar. 2002.

[20] R. P. Browne and P. D. McNicholas, “Model-based clustering, classification, and discriminant analysis of data with mixed type,” *J. Stat. Plan. Inference*, vol. 142, no. 11, pp. 2976–2984, 2012.

[21] L. Hunt and M. Jorgensen, “Clustering mixed data,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 4, pp. 352–361, 2011.

[22] W. J. Krzanowski, “The location model for mixtures of categorical and continuous variables,” *J. Classif.*, vol. 10, no. 1, pp. 25–49, Jan. 1993.

[23] C. M. Bishop and N. Nasrabadi, “Pattern Recognition and Machine Learning,” *Pattern Recognit.*, vol. 4, no. 4, p. 738, 2006.

[24] J. M. Joyce, “Kullback-Leibler Divergence,” in *International Encyclopedia of Statistical Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 720–722.

[25] D. Lam, M. Wei, and D. Wunsch, “Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning,” *IEEE Access*, vol. 3, pp. 1605–1616, 2015.

[26] A. Association, “2008 Alzheimer’s disease facts and figures,” *Alzheimer’s Dement.*, vol. 4, no. 2, pp. 110–133, 2008.

[27] S. Gauthier *et al.*, “Mild cognitive impairment,” *Lancet*, vol. 367, no. 9518. pp. 1262–1270, 2006.

[28] M. Castro and G. E. Smith, “Mild cognitive impairment and Alzheimer’s disease.,” in *APA handbook of clinical geropsychology, Vol. 2: Assessment, treatment, and issues of later life.*, Washington: American Psychological Association, 2015, pp. 173–207.

[29] Michael W. Weiner MD, “ADNI Alzheimer’s Disease Neuroimaging Initiative,” 2013. [Online]. Available: http://www.adni-info.org/. [Accessed: 18-Aug-2017].